A Mechanism Design Approach to Measure Awareness*

Diodato Ferraioli
DI, Università degli Studi di Salerno, Italy
dferraioli@unisa.it

Carmine Ventre
SCM, Teesside University, UK
C.Ventre@tees.ac.uk

Gabor Aranyi
SCM, Teesside University, UK
G.Aranyi@tees.ac.uk

Abstract

In this paper, we study protocols that allow to discern conscious and unconscious decisions of human beings; i.e., protocols that measure awareness. Consciousness is a central research theme in Neuroscience and AI, which remains, to date, an obscure phenomenon of human brains. Our starting point is a recent experiment, called *Post Decision Wagering* (PDW) (Persaud, McLeod, and Cowey 2007), that attempts to align experimenters' and subjects' objectives by leveraging financial incentives. We note a similarity with mechanism design, a research area which aims at the design of protocols that reconcile often divergent objectives through incentive-compatibility. We look at the issue of measuring awareness from this perspective. We abstract the setting underlying the PDW experiment and identify three factors that could make it ineffective: rationality, risk attitude and bias of subjects. Using mechanism design tools, we study the barrier between possibility and impossibility of incentive compatibility with respect to the aforementioned characteristics of subjects. We complete this study by showing how to use our mechanisms to potentially get a better understanding of consciousness.

Introduction

Can machines "feel" just like human beings do? This is a fundamental question in AI, which has attracted a number of contributions with often divergent answers, see, e.g., (Hofstadter 1979; Searle and others 1980). It is, however, a question that could be considered ill-posed since it is not exactly understood how humans "feel": it is, for example, not even clear how to establish whether decisions are taken consciously or not. To find an answer, a group of neuroscientists (Persaud, McLeod, and Cowey 2007) have recently introduced a protocol, named *Post-Decision Wagering* (PDW), as a means to "directly measure awareness". They consider three scenarios to motivate the effectiveness of PDW: an experiment run on a blindsight subject[1], artificial grammar task and Iowa gambling task. Due to the page limit, we next give only a succinct description of these scenarios, starting with the blindsight experiment. Before the experiment, the subject is instructed that in each trial either a visual stimulus would be presented in her blind field or no stimulus would be shown. After each possible exposure to the stimulus, the subject firstly has to make a decision on its absence/presence and then either wager high or low on the correctness of the decision. The other two experiments follow a similar pattern: an initial training phase in which subjects gain some "knowledge", a successive exposition to some sort of positive/negative *signal* and finally the subject's decision (with wagering on the correctness) about the "sign" of the signal. The positive (negative, resp.) signal corresponds to a string that follows (does not follow, resp.) a previously learned pattern in the artificial grammar task, and a pack of cards with positive (negative, resp.) expected gain in the Iowa gambling task. In all scenarios, the amount wagered is gained if the decision is correct and lost otherwise. The process is repeated, but the subject receives feedback on the accuracy of the decisions and the amount won/lost only after the last trial.

According to its designers, PDW is successful because participants with some awareness that their decision is correct will wager high. Then a failure to maximize cash rewards (i.e., correct answers do not always correspond to a high wager) would signal that some of the correct decisions were made unconsciously. However, some criticism on the effectiveness of PDW has been moved (see below).

This work stems from the observation that PDW connects very neatly to mechanism design. We want here to *theoretically* analyze the claim of (Persaud, McLeod, and Cowey 2007) and investigate the relation between mechanism design and awareness.

Our Contribution. PDW is designed to align the experimenter's objective of measuring awareness with the subject's (induced) objective of maximizing her financial gains. This very same approach is used in mechanism design. By using a game-theoretic terminology, we can call the information regarding the subject's awareness/unawareness of the stimulus as the *type* of the subject. The experimenter wants to design a mechanism that incentivizes the subject to reveal her type *truthfully*. The incentives correspond to the maximization of a utility function for money, with the type determining the expected outcome of the wagering lottery.

*The first two authors are ordered alphabetically, as common in Theoretical Computer Science literature. Third author only is listed by contribution.

[1]Brain lesions may lead to blindness on a portion of the visual field. Blindsight is the ability to respond to visual stimuli in the blind fields *even in absence of perception*, cf. (Celesia 2010).

This parallel seems to suggest that *rationality* is a fundamental assumption behind PDW just as it is for mechanism design. We then revisit the claim in (Persaud, McLeod, and Cowey 2007) and note that rationality might not be the only factor that influences the decision of the subject. We in fact identify other two factors of equal importance: *risk attitude* (how subjects evaluate probabilistic outcomes) and *bias* (external factors may induce the subjects to develop an expectation on the positivity/negativity of the signal in the next trial). Risk attitude and bias are these subjective features that have been often highlighted by psychologists and neuroscientists as unavoidable elements of the decision process (Green and Swets 1966). In particular, we prove that "PDW directly measures awareness" if and only if the subject is rational, risk-neutral (see below for a definition) and has no bias, thus formalizing the claim of (Persaud, McLeod, and Cowey 2007).

Given the limited scope of PDW's success, we then continue this study by investigating the existence of mechanisms with better properties. Towards this end, we consider direct revelation truthful mechanisms and therefore concentrate on mechanisms in which we have to incentivize the subject to report her true type. These declarations could be done through verbal reports (Cleeremans, Destrebecqz, and Boyer 1998; Shanks and St John 1994) or numerical confidence ratings (Dienes and Scott 2005). However, our interest is mainly theoretical and we do not imagine/propose to run these mechanisms in a real experimental setting. It is, however, worthwhile to study direct revelation mechanisms. Firstly, a good theory of mechanism design for awareness should provide fundamental insights into what can and cannot be done; by the revelation principle, this study can be restricted to direct mechanisms without loss of generality. Secondly, our results could serve as a natural starting point for the design of more practical protocols: the existence of a direct mechanism with the desired properties can foster neuroscience and AI experimenters to design new mechanisms with the same properties that only require indirect declaration of types and, thus, turn out to be more practical.

We begin by proving that there exists a truthful risk-independent mechanism, i.e., a mechanism which remains truthful for any possible risk attitude of the subject. As observed in (Clifford, Arabzadeh, and Harris 2008), truthfulness might be too weak a requirement in this context as the subject could make a declaration independent from the awareness itself that guarantees the same utility as the truthful declaration. We then look at the existence of strictly truthful mechanisms in which the utility is strictly maximized by truth-telling. We prove that there exists a strictly truthful mechanism which works under the assumption that the subject's risk attitude is known. We complement this positive result by showing that it is impossible to design a strictly truthful risk-independent mechanism. Our results highlight a trade-off between risk-independence and truthfulness, suggesting the use of one mechanism or the other depending on the priority between these requirements.

We then focus on the case in which the subject develops a bias during the several trials of the experiment. We prove that it is possible to induce the subject to truthfully report the presence/absence of bias. This can be useful for "filtering" experiments: only if the subject declares no bias, then we can adopt the above mechanisms to successfully measure the awareness. We also describe a realistic setting in which the rate of "no bias" reports can be carefully controlled.

Finally, we show how to use our results in the tentative to get a better grasp of how awareness arises.

Due to page limit, some proofs are omitted.

Related Works. Much of the related work in cognitive sciences criticizes PDW; these observations inform our work as we discuss below. Seth (2008) affirms that there is a fundamental flaw in the approach since "absence of evidence is not evidence of absence"; our mathematical model includes Seth's remark as a special case. Clifford, Arabzadeh, and Harris (2008) observe that for PDW "always bet high" maximizes the expected financial reward of the subject. They then note that PDW cannot measure awareness since this strategy is independent from the awareness. However, this is not entirely correct since "always bet high" is not the unique reward-maximizing strategy: "bet high when aware and low otherwise" achieves the same expected utility. Which reward-maximizing strategy will subjects play? This is debatable and motivates our quest for mechanisms with unique reward-maximizing strategies (i.e., strictly truthful mechanisms). According to (Schurger and Sher 2008), people exhibit loss aversion which could let them to wager low even when they are aware. This remark motivates our study of mechanisms that work for different levels of the subjects' risk attitude.

Related research topics include prediction markets and (market) scoring rules. While a similarity can be drawn with prediction markets (monetary value is given in both cases to private information/beliefs), the known manipulability of (market) scoring rules (Conitzer 2009) remains a fundamental difference. Proper scoring rules are instead quite similar to our truthful mechanisms. However, the main reason for which they do not correctly fit our setting is that they assume that the subject can always correctly describe the probability that the observed signal is positive. In our setting, instead, we consider also the case in which the granularity with which the subject describes her awareness is very gross-grained. With classical proper scoring rules, such as the spherical or Brier rule, it is not possible to incentivize subjects giving only gross-grained awareness descriptions to make the desired declarations. On the contrary, our payment rules can be seen as a proper scoring rule with the additional feature of working as desired also in this gross-grained setting.

Preliminaries

Let us start by abstracting the setting underlying the PDW experiments, in order to theoretically analyze the properties of PDW and of the other mechanisms developed in this work. The setting involves two entities, an *experimenter* E and a *subject* S. After a suitable learning phase, the experimenter sends a binary signal to the subject. Here, we assume that $s = 1$ corresponds to a positive signal (presence

of visual stimulus, a string following a pattern, a deck with positive gain) and $s = 0$ corresponds to a negative signal. Upon the dispatch, and the possible reception, of the signal s, the subject develops a *conscious* awareness about the value of received signal, that we call *awareness level* of S and denote with α_s. What values can α_s take? The literature in neuroscience suggests the existence of three states: the subject can be "sure that the signal is 1", "sure that the signal is 0", or "have no clue". We then reserve three different values of α_s for these states: 0 for awareness of signal 0, $1/2$ for unawareness, and 1 for awareness of signal 1. Other choices would be feasible (e.g., we can use $-1, 0, 1$). However, our choice can be easily extended in order to measure "strength of awareness", that is, to allow different and more fine-grained awareness levels, such as being "almost sure that the signal is 1". We indeed let $\alpha_s \in [0, 1]$, with the rationale that the closer to 0 (1, respectively) the awareness level is the more confident the subject is that the signal was negative (positive, respectively). We remark that the results in this paper hold whichever the granularity of the awareness level is.

We note that the awareness level depends on the received signal. This property allows us to meet the aforementioned Seth's critique (Seth 2008) that a subject cannot be aware of a negative signal ($s = 0$). This is in fact a special case of our model, wherein $\alpha_0 = 1/2$. Note also that we do not make any assumptions on how the awareness level is generated in the subject brain (i.e., how it depends on the signal). This process is actually unknown and its understanding represents the ultimate goal of this line of research.

Mechanism Design. The goal of PDW and of the other protocols developed in this work is to "measure" the awareness level. In other words, we want to design a mechanism M that allows to measure α_s, also called the subject's type in this context. For example, we might simply want to know whether $\alpha_s \neq 1/2$ (i.e., there is some awareness of the signal) or aim to a much more granular knowledge and request to know the exact value of α_s. The information that a mechanism wants to compute is modeled via a *choice function* that maps types into *outcomes*, more specifically, $f\colon [0, 1] \to X$, X denoting the set of allowed outcomes. For example, in the case in which we are only interested in distinguishing whether $\alpha_s = 1/2$ or not, X might simply be $\{0, 1\}$ and the choice function be $f(1/2) = 0$, $f(y) = 1$ for $y \in [0, 1/2) \cup (1/2, 1]$. Note that the outcome set X can contain monetary transfers (e.g., the wagers of PDW). Moreover, f might be a probabilistic function defining a probability distribution over X.

A subject of type $t \in [0, 1]$ evaluates an outcome $x \in X$ an amount $u_t(x)$ called *utility* (or *valuation*). Since there is no intrinsic utility in being aware or unaware, when X is a set of monetary transfers, the utility of a subject does not depend on her type, i.e., $u_t(x) = x$ for all $x \in X$. In other words, the subject evaluates an outcome exactly the amount of money she receives or loses.

A *mechanism* M defines a set of strategies A for the subject and an *output function* $g\colon A \to X$ specifying the output for each possible action chosen by the subject (as f, g might be a probabilistic function). A mechanism is said to *implement* in dominant strategies the choice function f if there exists $a \in A$, termed dominant strategy, s. t. (i) $u_t(g(a)) \geq u_t(g(a'))$ for all $a' \in A$ *and* (ii) $g(a) = f(t)$, t being the type of the subject. Additionally, a mechanism is said to *strictly implement* in dominant strategies the choice function f if a is the unique dominant strategy. (In case of probabilistic functions, the requirement (i) must naturally hold in expectation over the random coin tosses of g.) By the revelation principle (Nisan et al. 2007), we can restrict our attention w.l.o.g. to the class of *direct revelation* mechanisms in which the domain A of the output function is exactly the type domain $[0, 1]$. (Mechanisms for which $A \neq [0, 1]$ are instead called *indirect*.) In a direct mechanism, subjects can only declare types and strategizing equates to misreporting their true type. Therefore, the implementation concept of interest here is *truthfulness*, i.e., we are interested in mechanisms for which truth-telling is a dominant strategy. In the case in which truth-telling is the only dominant strategy, we say that the mechanism is *strictly truthful*.

We identify three factors that can influence the subject's utility and then fundamentally change her strategic behavior.

Rationality. A basic assumption in mechanism design and, more generally, social choice theory is that the subjects are rational, that is, they aim at their personal best outcome. By modeling personal preferences by means of money, this results in rational subjects being *utility maximizers*. When randomization is involved in the outcomes computation, it is usual to assume that subjects are *expected utility maximizers*.

Risk Attitude. The behavior of a subject can strongly deviate from the expected one when probabilistic outcomes (a.k.a., lotteries) are involved. In these cases, a fundamental role is played by the subject's *risk attitude*.

Two main definitions are known in economics literature to model people's attitude towards lotteries: expected utility theory, see, e.g., (Mas-Colell, Whinston, and Green 1995), and prospect theory (Kahneman and Tversky 1979). The former considers the subject's attitude to accept a bargain with uncertain reward rather than one with a certain, but possibly lower reward. It requires the existence of a *utility function for money* whose graph's curvature measures the risk attitude of the subject (wherein, concavity models risk aversion whilst convexity encodes risk seeking behavior). Prospect theory is, instead, a behavioral economic theory that states that people's risk attitude is based on the potential value of "losses" and "gains" rather than the final outcome.

Our definition of risk attitude is inspired by prospect theory and mainly motivated by the simplicity of the mechanisms considered in this work. A subject weights gains (e.g., a won PDW wager) with a factor $\rho_w \geq 0$ and losses (e.g., a lost PDW wager) with a factor $\rho_l > 0$. So, if the subject can obtain a reward R with probability p and a loss L with the remaining probability $1 - p$ then her utility will be $\rho_w \cdot p \cdot R - \rho_l \cdot (1 - p) \cdot L$. To simplify algebraic manipulations, we rewrite the utility as $\rho \cdot p \cdot R - (1 - p) \cdot L$,

where $\rho = \rho_w/\rho_l$. We call ρ the *risk attitude* of the subject. A mechanism is *risk-independent* if it is independent from the value of ρ (i.e., mechanism does not use the actual value of ρ). Typically, three different risk attitudes are considered. When an individual tends to avoid losses, i.e., $\rho < 1$, then we say she is *risk-averse* (or *risk-avoiding*). When the individual has $\rho = 1$, then we say she is *risk-neutral*. Finally, we say an individual is *risk-prone* (or *risk-seeking*) when $\rho > 1$.

Repeated Experiments and Bias. It should be noticed that the setting described above models only a single trial of a PDW experiment. This might look inadequate since the experiments use many trials. However, the utility of a subject for the whole experiment can be safely assumed to be the sum of the utility obtained at every single trial. (Note that we do not need a discount factor, since typically all the trials happen at around the same time.) Therefore, for our objectives it is enough to consider mechanisms for a sole trial.

Nevertheless, the fact that the trials are repeated adds some externalities we need to take into consideration. Indeed, a subject can use her knowledge of the previous trials to form a *bias* about the possible outcome, and, in turns, this bias can influence her decision. Note that, since no feedback and no money is given at any trial, there is no way for the subject to develop confidence or bias about (the correctness of) her awareness. Hence, we only consider bias as an expectation about the outcome of the next trial.

Different models, named *bias influence rules*, describe how the bias affects the subject's decision, e.g., (Chater and Oaksford 2008; Gal and Pfeffer 2008). Here, we assume a very general model: if outcome x has unbiased probability π of arising, then a subject with bias β will assign to x probability $p_\beta(\pi)$.

Mechanisms that Measure Awareness

We begin by formally defining the choice function that we want to implement. Our aim is to distinguish whether the signal has been perceived with some awareness or not, i.e. $\alpha_s \neq \frac{1}{2}$ or $\alpha_s = \frac{1}{2}$. For this reason, we need that $f(x) \neq f(1/2)$ for any $x \neq \frac{1}{2}$. Note that this defines an infinite family of functions, that we denote as *binary choice functions*.

Post-Decision Wagering. We now cast the PDW protocol in the mechanism design framework. The set of allowed outcomes is $X = \{-H, -L, L, H\}$ with $L < H$. We can interpret X as possible amount of money received (when positive) or to pay (when negative). The utility that a subject of type t associates to an outcome $x \in X$ is simply $u_t(x) = x$ and is therefore independent from t.

PDW is an indirect mechanism with strategy set $\{(b, w) \mid b \in \{0, 1\} \wedge w \in \{L, H\}\}$. Recall that in PDW if the experimenter's signal is s, then for an input action (b, w), the wager w is won if $b = s$ and lost otherwise, i.e., $g(b, w) = w$ if $b = s$ and $g(b, w) = -w$ otherwise. Therefore, PDW has a probabilistic output function g that on input the action (b, w) returns a lottery λ; λ assumes value w with the probability of the event "$b = s$" and value $-w$ with the probability of the event "$b \neq s$".

Next theorem shows mathematically when the "informal" claim of (Persaud, McLeod, and Cowey 2007) is true.

Theorem 1. *PDW implements a binary choice function if and only if the subject is rational, risk neutral and has no bias on the outcomes.*

Proof. Let us start by proving the "if" direction. Specifically, we prove that PDW implements the following specific binary choice function: $f(y) = \mathcal{D}^-(y)$ if $y \in [0, 1/2)$, $f(y) = \mathcal{U}$ if $y = 1/2$, and $f(y) = \mathcal{D}^+(y)$ if $y \in (1/2, 1]$, where $\mathcal{D}^+(y)$ ($\mathcal{D}^-(y)$, respectively) is a distribution which returns H with probability y ($1 - y$, respectively) and $-H$ with probability $1 - y$ (y, respectively); $\mathcal{U}$ is, instead, the uniform distribution over $\{L, -L\}$.

Let us describe the expected utility of the subject for the PDW mechanism. By definition of PDW, we have $u_{\alpha_s}(g(b, w)) = (-1)^{1-b} \cdot w \cdot \alpha_s + (-1)^b \cdot w \cdot (1 - \alpha_s)$, where we have used the assumptions that the subject is risk neutral and has no bias. Indeed, the subject's perceived probability that $s = 1$ is exactly α_s. Thus declaring $(1, w)$ gives a reward of w with such a probability and a loss of w with the remaining probability, whereas declaring 0 just inverts rewards with losses. Note that, by rationality, the subject aims to maximize $u_{\alpha_s}(g(b, w))$. It is not hard to check that $(1, H)$ is a dominant strategy and $g((1, H)) = f(\alpha_s)$ if $\alpha_s > \frac{1}{2}$; $(0, H)$ is a dominant strategy and $g((0, H)) = f(\alpha_s)$ if $\alpha_s < \frac{1}{2}$; and, finally, $(0, L)$ is a dominant strategy for $\alpha_s = \frac{1}{2}$ with $g((0, L)) = f(1/2)$.

As for the "only if" direction, we first observe that if the subject is not rational, then PDW has no way to distinguish if the action played falls in the rational responses or the irrational ones. Hence, if, for example, the subject bets low, then for the experimenter it is impossible to distinguish if this behavior derives from the unawareness of the signal or from the inability to maximize utility.

If the subject has a bias β on the possible signal s, then her utility will differ. In details, now $u_{\alpha_s}(g(b, w)) = (-1)^{1-b} \cdot w \cdot p_\beta(\alpha_s) + (-1)^r \cdot b \cdot (1 - p_\beta(\alpha_s))$, where $p_\beta(\alpha_s)$ represents the probability according to which the subject expects that $s = 1$ as function of bias and awareness. Thus, the subject will answer according to this updated utility function. However, PDW has no way to learn α_s from $p_\beta(\alpha_s)$ and, hence, it is not possible to distinguish if a low bet corresponds to $\alpha_s = 1/2$ or to $p_\beta(\alpha_s) = 1/2$ but $\alpha_s \neq 1/2$.

If the subject is risk-averse, that is $\rho_l > \rho_w$, then there is no outcome that turns out to be "reserved" to $\alpha_s = \frac{1}{2}$. Indeed, subject's utility is now $u_{\alpha_s}(g(b, w)) = (-1)^{1-b} \cdot w \cdot \alpha_s \cdot (\rho_w/\rho_l)^b - (-1)^b \cdot w \cdot (1 - \alpha_s) \cdot (\rho_w/\rho_l)^{1-b}$. Thus, if $\alpha_s \leq \frac{\rho_w}{\rho_w + \rho_l}$, then u_{α_s} is maximized by declaring $(0, H)$, whereas for $\frac{\rho_w}{\rho_w + \rho_l} < \alpha_s < \frac{1}{2}$, the utility is maximized by declaring $(0, L)$. Symmetrically, if $\alpha_s \geq \frac{\rho_l}{\rho_w + \rho_l}$, then u_{α_s} is maximized by declaring $(1, H)$, whereas for $\frac{1}{2} < \alpha_s < \frac{\rho_l}{\rho_w + \rho_l}$, the utility is maximized by declaring $(1, L)$.

Finally, if the subject is risk-seeking, i.e., $\rho_l < \rho_w$, then there is no outcome that is "reserved" to the case $\alpha_s \neq \frac{1}{2}$.

Indeed, the subject's utility for $\alpha_s = \frac{1}{2}$ is $u_{\alpha_s}(g(b,w)) = \frac{1}{2} \cdot w \cdot (\rho - 1)$, which is maximized by declaring $(1, H)$ or $(0, H)$. However, $(1, H)$ is also the dominant strategy when $\alpha_s = 1$ and $(0, H)$ is dominant when $\alpha_s = 0$. □

A Risk-Independent Truthful Mechanism. Let us now consider direct revelation mechanisms and therefore focus on truthful implementations of choice functions. These mechanisms allow to implement a choice function that is more fine-grained than that of Theorem 1, as we will be able to "measure" all the possible values of awareness level and not just distinguish awareness from unawareness.

Let us now define the mechanism. For a number $x \in [0,1]$, we let I_x be 1 if $x \geq 1/2$ and 0 otherwise. Our mechanism requires the subject to report a number $x \in [0,1]$. The desideratum is that x is exactly the subject's awareness level. To this aim, the mechanism does the following: If $I_x = s$ then the subject wins W otherwise she pays W, where W is any positive value. We call the mechanism D-PDW.

The utility that a subject of type t and risk attitude ρ attaches to the output of D-PDW, in input a declaration x, is then

$$u_t(x) = (-1)^{1-I_x} \cdot W \cdot \rho^{I_x} \cdot t + (-1)^{I_x} \cdot W \cdot \rho^{1-I_x} \cdot (1-t).$$

We want to prove that this mechanism gives the right incentives to the subject, that is, her utility is maximized when declaring her true type. Formally, we want to satisfy the following inequalities: $u_t(t) \geq u_t(x)$ for all $t, x \in [0,1]$. To prove this, we adopt the well-known cycle-monotonicity technique (Rochet 1987). We set up a weighted graph, called the *declaration graph*, associated to the algorithm above, with a vertex for each possible declaration, i.e., for each $x \in [0,1]$ and an arc between vertices x and y with weight $\delta(x,y) = u_x(x) - u_x(y)$, encoding the loss that a bidder whose type is x incurs into by declaring y. The following result relates the existence of negative edges in the declaration graph to the truthfulness of the algorithm.

Theorem 2. *(Rochet 1987) If the declaration graph associated to the algorithm does not have negative-weight edges then the algorithm is truthful. If the graph has only positive-weight edges then the algorithm is strictly truthful.*

By using this result, we can prove that D-PDW is truthful, regardless of the risk attitude of the subject.

Theorem 3. *D-PDW is a risk-independent truthful mechanism if the subject is rational and has no bias.*

Proof. Each edge of the graph associated to D-PDW has a non-negative weight. Indeed, for each $t \leq \frac{1}{2}$, $\delta(t,x) = (\rho W(1-t) - Wt) - (\rho W(1-t) - Wt) = 0$ if $x \leq \frac{1}{2}$ and $\delta(t,x) = (\rho W(1-t) - Wt) - (\rho Wt - W(1-t)) \geq 0$, otherwise. The case for $t > 1/2$ is symmetric and hence omitted. The theorem then follows from Theorem 2. □

Strictly Incentivizing Truth-telling. PDW and D-PDW are not strictly truthful, a requirement that appears to be useful in this context. We build upon D-PDW in order to guarantee this property at the cost of losing risk-independence.

The new mechanism also requires, as D-PDW, the subject to report a number $x \in [0,1]$; we let I_x denote the closest integer to x as above. Differently from D-PDW, rewards and losses here depend on the declaration. Specifically, the mechanism, that we call S-PDW, runs as follows: If $I_x = s$, then the subject wins $W(x)$, otherwise she pays $L(x)$, where $W(x) = \frac{3}{2}x^2 - x^3$ if $x \geq \frac{1}{2}$ and $W(x) = \frac{3}{2}(1-x)^2 - (1-x)^3$, otherwise, and $L(x) = x^3$ if $x \geq \frac{1}{2}$ and $L(x) = (1-x)^3$, otherwise.

The *utility* that a risk-neutral subject of type t attaches to the output of the algorithm above, in input a declaration x, is then $u_t(x) = W(x) \cdot t - L(x) \cdot (1-t)$ if $x \geq 1/2$, and $u_t(x) = W(x) \cdot (1-t) - L(x) \cdot t$, otherwise.

Theorem 4. *If the subject is rational, risk neutral and has no bias then S-PDW is strictly truthful. Additionally, the subject never experiences negative utility by truth-telling.*

Proof. Fix $t \geq \frac{1}{2}$. By looking at its prime derivative, $u_t(x)$ is increasing for $x \in [1/2, t]$ and decreasing for $x \in [t, 1]$. Hence, for any $x \geq 1/2$, with $x \neq t$, we have $\delta(t,x) = u_t(t) - u_t(x) > 0$. Similarly, it turns out that $u_t(x)$ is increasing for $x \in [0, 1/2]$. Hence, for $x < 1/2$, $\delta(t,x) = u_t(t) - u_t(x) > u_t(t) - u_t(1/2) = \frac{t^3}{2} - \frac{3t}{8} + \frac{1}{8} \geq 0$. The case for $t < \frac{1}{2}$ is symmetric and hence omitted.

Strict truthfulness then follows from Theorem 2 and the assumptions of rationality, risk neutrality and no bias.

The non-negativity of truth-telling subjects' utility follows from $u_t(t) \geq \frac{\max\{t^3, (1-t)^3\}}{2} \geq 0$ for any $t \in [0,1]$. □

Note that the values of functions W and L given above are only illustrative. There are indeed many other choices for these functions guaranteeing strict truthfulness.

The above mechanism works also in presence of a known risk attitude ρ by simply dividing the function W defined above by ρ. However, we might wonder if there is a mechanism that is able to strictly distinguish unawareness, i.e. $\alpha_s = \frac{1}{2}$, from awareness, i.e. $\alpha_s \neq \frac{1}{2}$, regardless of the specific value of ρ. Formally, we consider the question of whether it is possible to design a risk-independent mechanism that strictly implements a binary choice function. Next theorem shows that this is impossible.

Theorem 5. *There is no risk-independent mechanism that strictly implements a binary choice function.*

Proof. By the revelation principle, we can restrict our attention to direct mechanisms and strict truthfulness. According to (Blumrosen and Feldman 2006), since we would like to implement a binary choice function, we can assume that the subject has a restricted action space consisting only of $t_=$ corresponding to $\alpha_s = \frac{1}{2}$ and $t_\neq$ corresponding to $\alpha_s \neq \frac{1}{2}$. Since we are assuming risk independence, the only free parameters of the mechanism are the value 0/1 of the signal and the subject's declaration. Thus, for any subject's declaration $x \in \{t_=, t_\neq\}$, we can have only two possible outcomes whose realization depend on s. So, let $W_1(x)$ and $W_0(x)$ denote the pair of outcomes corresponding to the

declaration x. We can assume w.l.o.g. that the subject evaluates an outcome o exactly o. Thus, the expected utility of a subject with type t and risk attitude ρ in declaring x is $u_t(x) = W_1(x)\cdot\rho\cdot p_t + W_0(x)\cdot(1-p_t)$, if $W_1(x) \geq W_0(x)$, and $u_t(x) = W_1(x)\cdot p_t + W_0(x)\cdot\rho\cdot(1-p_t)$, otherwise, where p_t is the probability that the subject assigns to the signal being 1, i.e., $p_t = 1/2$ if $t = t_=$ and $p_t \neq 1/2$ otherwise. (We make no assumption on $W_1(\cdot)$ and $W_0(\cdot)$ and we leave to the bigger of the two the role of rewards.)

Assume that $W_1(t_=) \geq W_0(t_=)$ and $W_1(t_\neq) \geq W_0(t_\neq)$. Then, by strict truthfulness, it must be that $W_1(t_=)\cdot\rho\cdot\frac{1}{2} + W_0(t_=)\cdot\frac{1}{2} > W_1(t_\neq)\cdot\rho\cdot\frac{1}{2} + W_0(t_\neq)\cdot\frac{1}{2}$ and $W_1(t_\neq)\cdot\rho\cdot p_t + W_0(t_\neq)\cdot(1-p_t) > W_1(t_=)\cdot\rho\cdot p_t + W_0(t_=)\cdot(1-p_t)$. Let us set $\Delta_1 = W_1(t_=) - W_1(t_\neq)$ and $\Delta_0 = W_0(t_\neq) - W_0(t_=)$. Then, we must have $\Delta_1 > \frac{1}{\rho}\Delta_0$ and $\Delta_1 < \frac{1-p_t}{\rho\cdot p_t}\Delta_0$. Since, by risk independence, both Δ_1 and Δ_0 cannot depend on ρ, it is impossible to satisfy both these conditions for any possible value of ρ. E.g., consider $\Delta_0 > 0$ and $\rho \geq \rho_{\min}$, for some $\rho_{\min} > 0$. The above two inequalities yield $\rho^{-1} < \Delta = \frac{\Delta_1}{\Delta_0} < \frac{1-p_t}{\rho\cdot p_t}$. Now, it is either that $\Delta \geq 1/\rho_{\min}$ (in which case, the r.h.s. inequality is false for $\rho \geq \frac{1-p_t}{p_t}\rho_{\min}$) or $\Delta < 1/\rho_{\min}$ (l.h.s. inequality false for $\rho = \rho_{\min}$).

Similar arguments work if we invert the order between $W_1(t_=)$ and $W_0(t_=)$ or between $W_1(t_\neq)$ and $W_0(t_\neq)$. □

Dealing with Bias. We start by showing that the same ideas underlying the mechanisms described above can be reused for providing us an useful tool for handling bias on its own.

Lemma 1. *There is a truthful mechanism according to which a rational subject declares exactly her own bias. Moreover, if the risk attitude of the subject is known, the mechanism can be made strictly truthful.*

Proof. Consider D-PDW where x is the reported bias instead of the reported awareness level. We assume to run the mechanism before the actual trial (i.e., before that the experimenter sends the signal to the subject). As mentioned above, we delay the payment to just after the experiment. Since the bias is reported before the signal is sent and, hence, before the awareness level is generated, the subject's utility depends only on her real bias t, possibly different from the declared bias x, and not on the awareness level. That is $u_t(x) = (-1)^{1-I_x}Wt + (-1)^{I_x}W(1-t)$. The result then follows from Theorem 3. By using S-PDW in place of D-PDW, Theorem 4 proves strict truthfulness. □

For sake of readability let us distinguish the application of the D-PDW mechanism for the bias from the one for the awareness level, by naming the former as BD-PDW and the latter as AD-PDW. It looks then natural to sequentially compose BD-PDW and AD-PDW to have successful experiments also in presence of bias. Unfortunately, this does not work without further assumptions. Indeed, as described above, the utility that the subject attaches to AD-PDW in presence of bias β, does not depend only on the awareness level α_s, but it is a function $p_\beta(\alpha_s)$ of both the awareness level and the bias. Thus, when we assume bias is present, AD-PDW cannot be able to learn the awareness level, but just the output $p_\beta(\alpha_s)$ of the bias influence rule, from which it may be not possible to know the awareness level.

For example, if the bias influence rule is given by an average of awareness level α_s and bias β weighted by weights w_{α_s} and w_β, i.e., $p_\beta(\alpha_s) = \frac{w_{\alpha_s}\alpha_s + w_\beta\beta}{w_{\alpha_s}+w_\beta}$, then the composition of the two mechanisms only allows to known β and $\frac{w_{\alpha_s}\alpha_s + w_\beta\beta}{w_{\alpha_s}+w_\beta}$. But this is not sufficient to extract α_s. However, if the experimenter knows how the subject mixes her prior knowledge with the awareness, then we can compose the mechanisms, and we have the following theorem.

Theorem 6. *If the bias influence rule is known, then there is a truthful mechanism for measuring awareness of rational subjects. Moreover, if also the risk attitude is known, the mechanism can be made strictly truthful.*

The theorem above replaces the assumption of no bias with the knowledge of the bias influence rule. Thus, it is natural to ask whether we can weaken the latter requirement or completely remove it.

Our idea is to have some control on the external environment so to "guide" the subject's bias towards situations that we know how to handle. In the following, we give a result following this line of thought.

Assume that the bias is simply a measure of the unpredictability of the next output. Formally, consider the binary string representing the values 0/1 of the signal in the previous trials of the experiment. Then, we assume the bias depends only on some randomness measure of this string (there are a plenty of them, some of theoretical flavor, such as the Kolmogorov complexity, and some using statistics, such as frequency tests). Thus, we expect that no bias exists when the input string is evaluated as being random, whereas bias is likely to arise when it appears to follow some pattern, and thus it is likely for the subject to predict the next output.

However, it is possible to prove that a string taken uniformly at random will be evaluated as random regardless of the randomness measure we are considering. More specifically, by a simple counting argument, any randomly selected string of n bit will have Kolmogorov complexity at least $n-c$, for some $c < n$, with probability at most $1-2^{-c}$. Since any randomness measure must be bounded from below by Kolmogorov complexity (i.e. strings with high Kolmogorov complexity should be declared random by any randomness test) then the result can be extended as desired.

Theorem 7. *If the experimenter chooses the signal uniformly at random and the subject bias depends only on some randomness measure on the outputs of previous trials, then there is a truthful mechanism for measuring awareness in a rational subject that works for almost any experiment with high probability. If also the risk attitude of the subject is known, the mechanism can be made strictly truthful.*

Learning How the Awareness Arises

Assume there is a discrete probability distribution D_s such that, ceteris paribus, the awareness level of the subject when she receives the signal s is α with probability $D_s(\alpha)$. Can we then learn D_s? That is, can we learn how often (and how

strongly) the subject is aware of the received signal? And how many trials do we need in order to be confident that the computed distribution closely approximates D_s?

Interestingly, a simple application of the Hoeffding bound shows that with a relatively small number of trials our proposed mechanisms can compute a distribution that is "close" to D_s.

Theorem 8. *If the true awareness can be learned at each experiment, then, for each $\varepsilon > 0$ we can learn a distribution D'_s that has total variation distance at most ε from D_s, by repeating the experiment $\tilde{O}\left(|\mathrm{Supp}(D_s)|^2/\varepsilon^2\right)$ times, where $\mathrm{Supp}(D_s)$ denotes the support of D_s.*

Conclusions

We proved that PDW is a good tool to measure awareness, if and only if the experiment is run on fully rational, risk neutral subjects who have no bias. Inspired by PDW, we provide new mechanisms that allow to measure awareness under different hypotheses on subject's rationality, risk attitude and bias. These results are proved via a novel connection between awareness and mechanism design/game theory (and then, in a larger sense, rationality). Moreover, our study improves the state of the art in neuroscience by recognizing and studying factors that can influence the subjects' wagers.

Our mechanisms require the subject to behave rationally. While the simplicity of our mechanisms can be advocated for the validity of such an assumption, this represents a practical limitation inherited by PDW. For example, when it is not possible to infer or train people in the lab to be utility maximizers then PDW and any of our mechanisms become ineffective. It is an interesting open problem to come up with ideas of experiments allowing, at the very least, a loose control on the rationality (and/or risk attitude, bias, etc.) of subjects under experimentation. This will make practically effective the purely theoretical contribution of our theorems. Nevertheless, we believe that our theoretical results have nontrivial implications and can inspire the design of new protocols fitting real experimental settings.

The general question of interest here seems to be the relation between rationality and consciousness. Where does rationality stand in the frontier between consciousness and unconsciousness? Such a question should consider claims about our unconscious rationality (Beck et al. 2008).

Acknowledgments. This work was partially funded by the European Commission under grant agreement CEEDs (FP7-ICT-258749). The authors wish to thank Marc Cavazza for introducing PDW and Valentina Bosco for pointing to relevant literature in Psychology.

References

Beck, J. M.; Ma, W. J.; Kiani, R.; Hanks, T.; Churchland, A. K.; Roitman, J.; Shadlen, M. N.; Latham, P. E.; and Pouget, A. 2008. Probabilistic population codes for bayesian decision making. *Neuron* 60(6):1142–1152.

Blumrosen, L., and Feldman, M. 2006. Implementation with a bounded action space. In *Proceedings of the 7th ACM conference on Electronic commerce*, 62–71. ACM.

Celesia, G. G. 2010. Visual perception and awareness: A modular system. *Journal of Psychophysiology* 24(2):62.

Chater, N., and Oaksford, M. 2008. *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press.

Cleeremans, A.; Destrebecqz, A.; and Boyer, M. 1998. Implicit learning: News from the front. *Trends in cognitive sciences* 2(10):406–416.

Clifford, C. W.; Arabzadeh, E.; and Harris, J. A. 2008. Getting technical about awareness. *Trends in cognitive sciences* 12(2):54–58.

Conitzer, V. 2009. Prediction markets, mechanism design, and cooperative game theory. In *Proc. of UAI*, 101–108.

Dienes, Z., and Scott, R. 2005. Measuring unconscious knowledge: Distinguishing structural knowledge and judgment knowledge. *Psychological research* 69(5-6):338–351.

Gal, Y., and Pfeffer, A. 2008. Networks of influence diagrams: A formalism for representing agents beliefs and decision-making processes. *Journal of Artificial Intelligence Research* 33(1):109–147.

Green, D. M., and Swets, J. A. 1966. *Signal detection theory and psychophysics*, volume 1. Wiley New York.

Hofstadter, D. R. 1979. *Godel, Escher, Bach: An Eternal Golden Braid*. New York, NY, USA: Basic Books, Inc.

Kahneman, D., and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 263–291.

Mas-Colell, A.; Whinston, M. D.; and Green, J. R. 1995. *Microeconomic Theory*. Oxford University Press.

Nisan, N.; Roughgarden, T.; Tardos, E.; and Vazirani, V. V. 2007. *Algorithmic Game Theory*. New York, NY, USA: Cambridge University Press.

Persaud, N.; McLeod, P.; and Cowey, A. 2007. Post-decision wagering objectively measures awareness. *Nature neuroscience* 10(2):257–261.

Rochet, J.-C. 1987. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of Mathematical Economics* 16(2):191–200.

Schurger, A., and Sher, S. 2008. Awareness, loss aversion, and post-decision wagering. *Trends in Cognitive Sciences* 12(6):209–210.

Searle, J. R., et al. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3(3):417–457.

Seth, A. K. 2008. Post-decision wagering measures metacognitive content, not sensory consciousness. *Consciousness and cognition* 17(3):981–983.

Shanks, D. R., and St John, M. F. 1994. Characteristics of dissociable human learning systems. *Behavioral and brain sciences* 17(03):367–395.